

# **FLASH MEMORY WITH ENHANCED PROGRAM AND ERASE COUPLING AND PROCESS OF FABRICATING THE SAME**

## **Background of the Invention**

### **Field of Invention**

This invention pertains generally to semiconductor memory devices and, more particularly, to a self-aligned split-gate flash memory and process of fabricating the same.

### **Related Art**

Nonvolatile memory is currently available in several forms, including electrically programmable read only memory (EPROM), electrically erasable programmable read only memory (EEPROM), and flash EEPROM. Flash memory has been widely used for high volume data storage in devices such as memory cards, personal digital assistants (PDA's), cellular phones, and MP3 players. Such applications require high density memory, with smaller cell size and reduced cost of manufacture.

In general, there are two basic types of nonvolatile memory cell structures: stack-gate and split-gate. The stack-gate flash memory cell usually has a bit line contact, a source region, a floating gate, and a control gate, with the control gate being positioned directly above the floating gate. In a split-gate cell the control gate is still positioned above the floating gate, but it is offset laterally from it. The fabrication process for a stack-gate cell is generally simpler than that for a split-gate cell. However, a stack-gate cell has an over-erase problem which a split-gate cell does not have. This problem is commonly addressed by maintaining the threshold voltage of the cell in a range of about 1.0-2.0 volts after an erase cycle, which adds complexity to the circuit design.

Although a split-gate memory cell has no over erase problem, it generally includes an additional gate known as a select gate. Such cells are typically

5 fabricated in double-poly or triple-poly processes which involve relatively complex processing steps. In addition, split-gate cells are generally larger than stack-gate cells. Nevertheless, because of the relatively simple circuit design which is possible when there is no over-erase problem, split-gate cells are used widely, particularly in embedded nonvolatile memory applications.

10 Figure 1 illustrates a split-gate, self-aligned flash memory cell 16 which is described in detail in U.S. Patent 6,091,104. This cell has a silicon substrate 17, with drain and source regions 18, 19 formed in a channel region 21. A floating gate 22 and a control gate 23 are formed above the channel region, with a gate oxide 24 between the substrate and the floating gate, and a dielectric film 26 between the floating gate and the control gate. A select gate 27 is formed to one side of the floating gate and control gate, with an oxide layer 28 between the select gate and the substrate, a dielectric film 29  
15 between the select gate and the control gate, and another dielectric film 30 between select gate and the floating gate.

20 In the program mode, control gate 23 is biased at a high positive voltage (e.g., 10-12 volts), the source at another high positive voltage (e.g., 6-8 volts), select gate at a lower positive voltage (e.g., 1-2 volts), and the drain at 0 volts. This establishes an electric field across off-gate channel region 25 between the floating gate and select gate, which initiates hot electron injection, with electrons migrating from the channel region to the floating gate.

25 In one erase mode, a negative voltage of about -10 volts is applied to control gate, a positive voltage of about 7 volts is applied to the select gate, and the source and drain are left floating. The strong electric field thus created across inter-poly dielectric film 30 initiates Fowler-Nordheim tunneling, with electrons migrating from the floating gate to the select gate. In another erase mode, a  
30 negative voltage of about -10 volts is applied to control gate, a positive voltage of about 7 volts is applied to the source, and the drain and select gate are left floating. The strong electric field thus created across gate oxide 24 initiates Fowler-Nordheim tunneling, with electrons migrating from the floating gate to the source.

35

As fabrication processes improve and geometries get smaller, e.g. tens of nanometers, it is difficult to form a high-voltage coupling ratio which is sufficient

for program and erase operations while keeping cell size small and meeting stringent reliability requirements such as 10-year data retention and 1,000,000 cycling operations between failures.

5     **Objects and Summary of the Invention**

It is in general an object of the invention to provide a new and improved semiconductor device and process for fabricating the same.

10     Another object of the invention is to provide a semiconductor device and process of the above character which overcome the limitations and disadvantages of the prior art.

15     These and other objects are achieved in accordance with the invention by providing a self-aligned split-gate flash memory cell array and process of fabrication in which erase and select gates are positioned on opposite sides of stacked floating and control gates, with source regions in the substrate beneath the erase gates, bit line diffusions which are partially overlapped by select gates at the ends of the rows of the cells. The floating and control gates are self-aligned with each other, and the erase and select gates are split from but  
20     self-aligned with the stacked gates.

25     An erase path is formed between the floating gate and the channel region beneath it, and electron migration from the floating gate to the channel region is produced by Fowler-Nordheim tunneling due to high voltage coupling between the control, select and erase gates and the floating gate. Programming is done by hot carrier injection from an off-gate channel region to the floating gate due to high voltage coupling between the control gate, erase gate and source region and the floating gate. The array is biased so that all of the memory cells in it can be erased simultaneously, while programming is bit selectable.

30

**Brief Description of the Drawings**

Figure 1 is a cross-sectional view of a self-aligned split-gate flash memory cell of the prior art.

35     Figure 2 is a cross-sectional view, taken along line 2 – 2 in Figure 3, of one embodiment of a self-aligned split-gate flash memory cell array incorporating the invention.

Figure 3 is a top plan view of the embodiment of Figure 2.

5      Figures 4A – 4E are schematic cross-sectional views illustrating the steps in one embodiment of a process for fabricating a self-aligned split-gate flash memory cell array in accordance with the invention.

10      Figures 5 is a circuit diagram of a small memory array as in the embodiment of Figure 2, showing exemplary bias conditions for erase, program and read operations.

### **Detailed Description**

15      As illustrated in Figures 2 and 3, the memory includes an array of self-aligned split-gate flash memory cells 36, each of which has stacked floating and control gates 37, 38. The control gates are positioned above and in vertical alignment with the floating gates.

20      Erase gates 43 and select gates 44 are positioned on opposite sides of the stacked gates. In the embodiment illustrated, there are two cells in each row, and erase gate 43 is positioned between the stacked gates, with select gates 44 on the outer sides of the stacked gates at the ends of the rows. The erase gates and select gates thus alternate between successive pairs of stacked gates, with the gates at the ends of the rows being select gates regardless of the number of cells in the rows.

25      An oxide layer 40 separates the floating gates from the substrate 41 below, and an inter-poly dielectric 42 is provided between the floating gates and the control gates. A dielectric 47 extends between the side walls of the floating and control gates and the erase and select gates.

30      A common source diffusion 49 is formed in an N-type well 52 in the upper portion of substrate 41 beneath erase gate 42, with the inner edge portions of floating gates 37 overlapping the outer edge portions of the source diffusion. Bit line diffusions 50 are formed in the P-type well at the ends of the rows and lie partially  
35      beneath the outer edge portions of select gates 44. Dielectric layers 53, 54 separate the erase gates and the select gates from the substrate.

A bit line 57 is connected to the bit line diffusions at the two ends of the row by bit line contacts 46.

5 As best seen in Figure 3, isolation regions 56 are formed in the substrate between rows of cells, and control gates 38 cross over floating gates 37 and the isolation regions. Erase gates 43 and select gates 44 extend in a direction parallel to the control gates. Bit lines 57 are perpendicular to the erase, select and control gates, and cross over the bit lines contact 46, erase gate, select gates, and stacked gates in each row of the array. Erase paths extend from the  
10 floating gates through tunnel oxides 40 to the channel regions beneath.

A preferred process of fabricating the embodiment of Figures 2 - 3 is illustrated in Figures 4A - 4E. In this process, oxide layer 40 is thermally grown to a thickness of about 60Å to 120Å on a monocrystalline silicon substrate which, in the  
15 embodiment illustrated, is in the form of a P-type substrate 41 in which P-type well 52 is formed. Alternatively, if desired, an N-type well can be formed in the P-type substrate, in which case the P-type well will be formed in the N-type well.

A conductive layer 62 of polysilicon or amorphous silicon (poly-1) is deposited on the thermal oxide to a thickness on the order of 300Å to 1500Å, and etched to  
20 form rows of stripes of silicon which coincide with the rows of cells to be formed on the substrate. The inter-poly dielectric layer 42 is formed on the silicon. The silicon is preferably doped with phosphorus, arsenic or boron to a level on the order of  $10^{17}$  to  $10^{20}$  per  $\text{cm}^3$ . The doping can be done *in-situ* during deposition  
25 of the silicon or by ion implantation either directly into the silicon or through the dielectric 42 above it. The inter-poly dielectric can be either a pure oxide or a combination of oxide, nitride and oxide (ONO), and in the embodiment illustrated, it consists of a lower oxide layer having a thickness on the order of 30Å - 100Å, a central nitride layer having a thickness on the order of 60Å - 200Å,  
30 and an upper oxide layer having a thickness on the order of 30Å - 100Å.

A second layer 63 of polysilicon (poly-2) is deposited on dielectric film 42. This layer has a thickness on the order of 1500Å - 3500Å, and is doped with phosphorus, arsenic or boron to a level on the order of  $10^{20}$  to  $10^{21}$  per  $\text{cm}^3$ . A  
35 CVD oxide or nitride layer 66 having a thickness on the order of 300Å - 1000Å is deposited on the poly-2 layer, and is used as a mask to prevent the poly-2 material from etching away during subsequent dry etching steps.

A photolithographic mask 67 is formed over layer 66 to define the control gates, and the unmasked portions of that layer and poly-2 layer 63 are etched away anisotropically, leaving only the portions of the poly-2 which form the control gates 38. The control gates extend in a direction perpendicular to the rows of silicon in the poly-1 layer. The exposed portions of the inter-poly dielectric 42 and the underlying portions of the poly-1 layer 62 are then etched away anisotropically to form the floating gates 37, as illustrated in Figure 4B. Thereafter, common source diffusion region 49 is formed in the substrate between the stack gates by ion implantation using dopants such as P<sup>31</sup> or As<sup>75</sup>.

Following ion implantation, dielectric 47 is formed on the sidewalls of control and floating gates, dielectric 53 is formed on common source diffusion 49, dielectric 54 is formed on silicon substrate, and a conductive (poly-3) layer 59 is deposited over the entire wafer, as shown in Figure 4C. Dielectric films 47, 53, 54 can be either pure oxide or the combination of oxide, nitride and oxide (ONO), and in the embodiment illustrated, they each consist of a lower oxide layer having a thickness on the order of 30Å - 100Å, a central nitride layer having a thickness on the order of 60Å - 300Å, and an upper oxide layer having a thickness on the order of 30Å - 100Å. The poly-3 layer is typically doped polysilicon or polycide, and is deposited to a thickness on the order of 1500Å - 3000Å.

The poly-3 layer is then etched anisotropically to form select gates 44 and erase gate 43, as illustrated in Figure 4D. Being formed in this manner, the select and erase gates are self-aligned with and parallel to the control gates. N-type dopants such as P<sup>31</sup> or As<sup>75</sup> are implanted into P-well 52 to form the bit line diffusions 50.

Thereafter, a glass material 60 such as phosphosilicate glass (PSG) or borophosphosilicate glass (BPSG) is deposited across the entire wafer, then etched to form openings for bit line contacts 46, as shown in Figure 4E. Finally, a metal layer is deposited over the glass and patterned to form bit lines 57 and bit line contacts 46.

Operation and use of the memory array can now be described with reference to Figure 5 where exemplary bias voltages for erase (ERS), program (PGM) and

read (RD) operations are shown next to the terminals of the array. In this example, memory cell  $C_{1n}$  is selected. This cell is located at the intersection of control gate  $CG_1$  and bit line  $BL_n$ , and is circled on the drawing for ease of location. All of the other memory cells in the array are unselected.

5

During an erase operation, electrons are forced to tunnel from the floating gate to the channel region beneath it, leaving positive ions in the floating gate. This can be done with either of two different bias conditions. In the first erase mode (ERS1), control gates  $CG_0$ - $CG_1$  are biased at a level on the order of -10 to -13 volts, select gates  $SG_0$ - $SG_1$  and erase gate  $EG_0$  are biased at -6 to -13 volts, and the bit line, common source and P-well are biased at 0 volts. In the second erase mode (ERS2), the control, erase and select gates are biased at 0 volts, bit line and common source are floating, and the P-well is biased at 10 to 13 volts.

15 With these bias conditions, most of the voltages applied to the control gate, the select gate and the erase gate appear across the tunnel oxide under the floating gate. That triggers Fowler-Nordheim tunneling, with electrons tunneling from the floating gate to the channel region below. As the floating gate becomes more positively charged, the threshold voltage of the memory cell, which is preferably  
20 on the order of -2 to -5 volts in this embodiment, becomes lower. This results in an inversion layer in the channel beneath the floating gate when the control gate is biased at 0-1.5 volts. Therefore, the memory cell goes into the conductive state (logic "1") after the erase operation.

25 In the first erase mode, the control gates, the select gates and the erase gates in the unselected memory cells are biased at 0 volts, so there is no Fowler-Nordheim tunneling in them. In the second erase mode, in the unselected memory cells, the control gates are biased at 5 to 10 volts, the select gates and the erase gates are biased at 0 to 5 volts, the voltage across the  
30 tunnel oxide is greatly reduced and there is no Fowler-Nordheim tunneling.

With the control gate, select gate and erase gate surrounding the floating gate or cathode electrode, the capacitance between the gates is relatively large, and high-voltage coupling from these gates to the floating gate is substantially enhanced. That significantly reduces the voltage required for Fowler-Nordheim  
35 tunneling, and also makes it possible to use a thicker tunnel oxide while still maintaining sufficient electron tunneling.

During a program operation, the control gate of the selected memory cell  $C_{1n}$  is biased to a level of 9 to 11 volts, the select gate  $SG_1$  is biased at 1 - 2 volts, the erase gates  $EG_0$  is biased at 4 - 8 volts, the common source  $CS_0$  is biased at 4 - 8 volts, the bit line is biased at 0 volts, and P-well is held at 0 volts. With these bias conditions, most of the voltage between the common source and the bit line appears across the mid-channel region between select gate  $SG_1$  and the floating gate of the selected cell  $C_{1n}$ , resulting in a high electric field in that region. In addition, since the floating gate is coupled to a high voltage from the control gate, erase gate and common source, a strong vertical electric field is established across the oxide between the mid-channel region and the floating gate. When electrons flow from the bit line to the common source, they are accelerated by the electric field across the mid-channel region, and some of them become heated. Some of the hot electrons get accelerated by the vertical field, which causes them to overcome the energy barrier of the oxide (about 3.1 eV) and inject into the floating gate.

At the end of the program operation, the floating gate is negatively charged, and the threshold voltage of the memory cell, which preferably is on the order of 3 - 5 volts, becomes higher. Thus, the memory cell is turned off when the control gate is biased at 0 volts during a read operation. Following a program operation, the memory cell goes into a non-conductive state (logic "0").

In the unselected memory cells  $C_{1(n-1)}$  and  $C_{1(n+1)}$  which share the same control gate with the selected cell  $C_{1n}$ , the bit line is biased at 3 volts, the select gate  $SG_1$  is at 1 - 2 volts, and the control gate is at 9 - 11 volts. Thus, select transistors  $S_{1(n-1)}$  and  $S_{1(n+1)}$  are turned off, and there is no mid-channel hot carrier injection in cells  $C_{1(n-1)}$  and  $C_{1(n+1)}$ . For the unselected memory cells such as  $C_{0n}$  and  $C_{2n}$  in the selected bit line, the control gates and select gates are biased at 0 volts, the bit line is biased at 0 volts, which minimizes the mid-channel hot carrier injection, and the floating gate charges are unchanged.

The control gates, the erase gates, and the common source nodes surround the floating gates in a manner which provides relatively large inter-gate and common source to floating gate capacitances for high-voltage coupling during program operation.



In the read mode, the control gate of the selected memory cell  $C_{1n}$  is biased at 0 - 1.5 volts, the common source is biased to 0 volt, 1 - 3 volts is applied to the bit line,  $V_{cc}$  is applied to the select gate  $SG_1$ , and 0 volts is applied to the erase gate  $EG_0$ . When the memory cell is erased, the read shows a conductive state because the channel of selected cell is turned on. Thus, a logic "1" is returned by the sense amplifier. When the memory cell is programmed, the read shows a non-conductive state because the channel of the selected cell is turned off, and hence the sense amplifier returns logic "0".

In the unselected memory cells  $C_{0n}$  and  $C_{2n}$  in the selected bit line, the control gates and select gates are biased at 0 volts, and there is no current flow between the bit line and common source nodes. In the unselected memory cells  $C_{1(n-1)}$  and  $C_{1(n+1)}$ , both the bit line and common source nodes are biased at 0 volts, and there is no current flow between the bit line and the common source nodes.

The invention has a number of important features and advantages. It provides a self-aligned split-gate flash memory cell array which has enhanced high voltage coupling from the control gate, select gate and erase gate to the floating gate for erase operations, and from the control gate, erase gate and common source to the floating gate for program operations. Each cell has control and floating gates which are stacked and self-aligned with each other, and select and erase gates which are split from but self-aligned with the other two. With this enhanced coupling, the high voltage required for erase operation can be greatly reduced. Moreover, the memory cells can be made smaller, and the cell density can be greater than in memory structures heretofore provided.

The erase gates have heavily doped diffusion regions beneath them, but the select gates do not. The erase path is from the floating gate to the channel region beneath it, with Fowler-Nordheim tunneling, the program path is from the off-gate channel region to the floating gate near it, using hot carrier injection, and the array is biased so that all of the memory cells in the array can be erased simultaneously, while programming is bit selectable.

It is apparent from the foregoing that a new and improved self-aligned split-gate flash memory and process of fabrication have been provided. While only certain presently preferred embodiment has been described in detail, as will be apparent to those familiar with the art, certain changes and modifications can be

made without departing from the scope of the invention as defined by the following claims.